Introduction to Bayesian Optimization

Javier González

Masterclass, 7-February, 2107 @Lancaster University

Big picture

"Civilization advances by extending the number of important operations which we can perform without thinking of them."
(Alfred North Whitehead)

We are interested on optimizing data science pipelines:

- Automatic model configuration.
- Automate the design of physical experiments.

Agenda of the day

- **9:00-11:00, Introduction to Bayesian Optimization**:
  - What is BayesOpt and why it works?
  - Relevant things to know.

- **11:30-13:00, Connections, extensions and applications**:
  - Extensions to multi-task problems, constrained domains, early-stopping, high dimensions.
  - Connections to Armed bandits and ABC.
  - An applications in genetics.

- **14:00-16:00, GPyOpt LAB!**: Bring your own problem!

- **16:30-15:30, Hot topics current challenges**:
  - Parallelization.
  - Non-myopic methods
  - Interactive Bayesian Optimization.

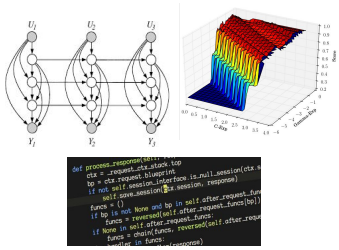
Section I: Introduction to Bayesian Optimization

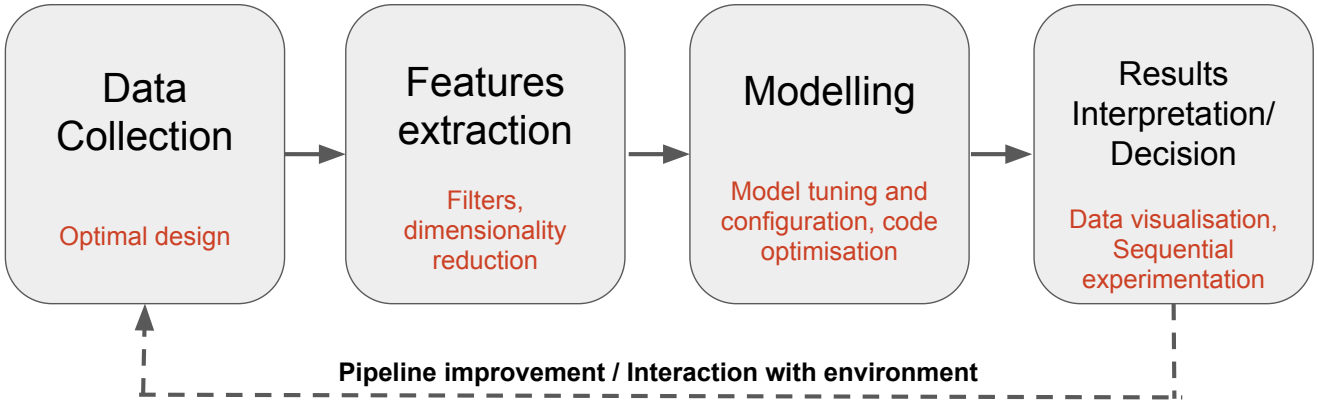
- What is BayesOpt and why it works?
- Relevant things to know.

Data Science pipeline/Autonomous System

Challenges and needs for automation





Experimental Design - Uncertainty Quantification

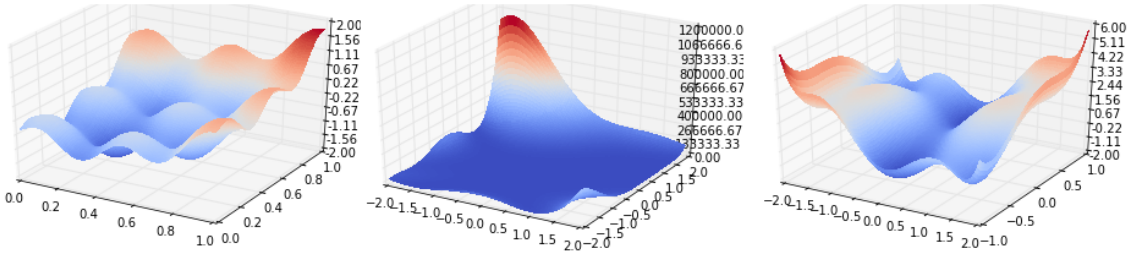
Can we automate/simplify the process of designing complex experiments?

Emulator - Simulator - Physical system

Global optimization

Consider a 'well behaved' function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subseteq \mathbb{R}^D$ is a bounded domain.

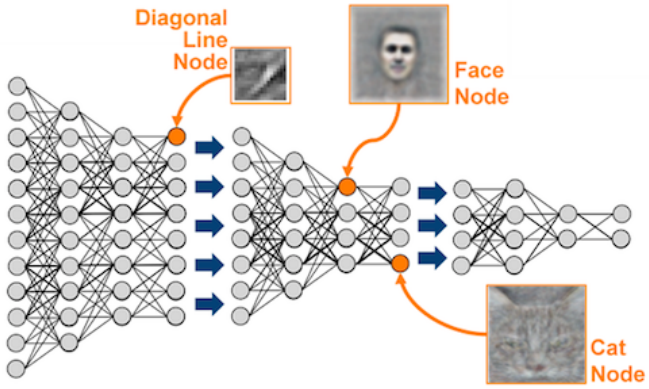
$$x_M = \arg\min_{x \in \mathcal{X}} f(x).$$



- f is explicitly unknown and multimodal.
- Evaluations of f may be perturbed.
- Evaluations of f are expensive.

Expensive functions, who doesn't have one?

Parameter tuning in ML algorithms.



- Number of layers/units per layer
- Weight penalties
- Learning rates, etc.

Figure source: http://theanalyticsstore.com/deep-learning

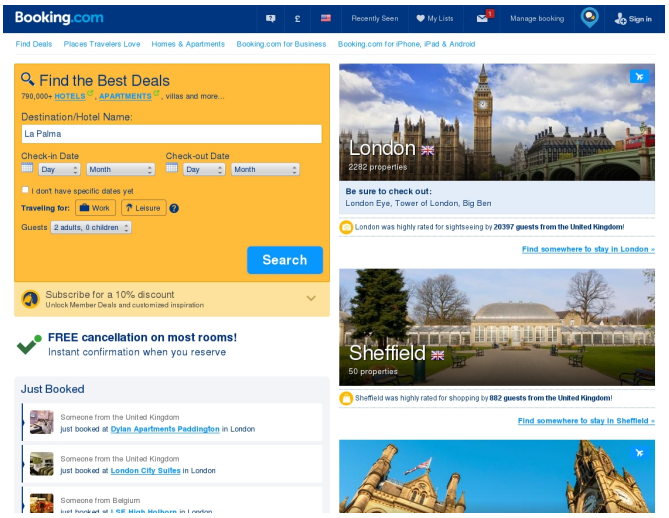
Expensive functions, who doesn't have one?

Active Path Finding in Middle Level

Optimise the location of a sequence of waypoints in a map to navigate from a location to a destination.

Expensive functions, who doesn't have one?

Tuning websites with A/B testing

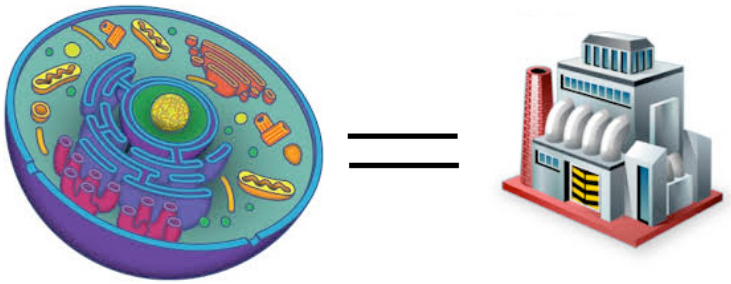


Optimize the web design to maximize sign-ups, downloads, purchases, etc.

Expensive functions, who doesn't have one?

[González, Lonworth, James and Lawrence, NIPS workshops 2014, 2015]

Design of experiments: gene optimization



- Use mammalian cells to make protein products.
- Control the ability of the cell-factory to use synthetic DNA.

Optimize genes (ATTGGTUGA...) to best enable the cell-factory to operate most efficiently.

Expensive functions, who doesn't have one?

Many other problems:

- Robotics, control, reinforcement learning.
- Scheduling, planning
- compilers, hardware, software?
- Intractable likelihoods.

What to do?

Option 1: Use previous knowledge

To select the parameters at hand. Perhaps not very scientific but still in use...

What to do?

Option 2: Grid search?

If f is L-Lipschitz continuous and we are in a noise-free domain to guarantee that we propose some $x_{M,n}$ such that

$$f(x_M) - f(x_{M,n}) \leq \epsilon$$

we need to evaluate f on a D-dimensional unit hypercube:

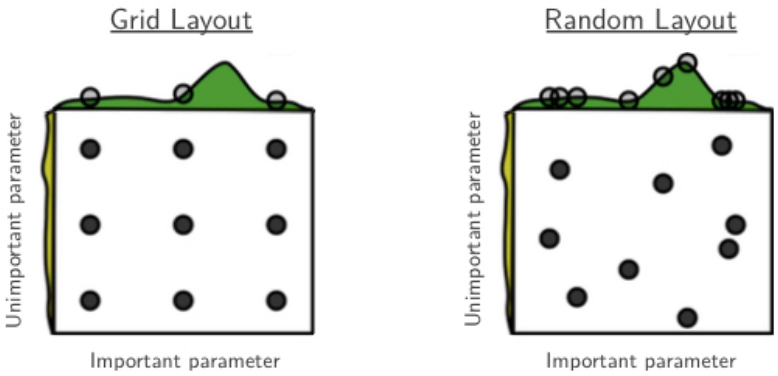
$$(L/\epsilon)^D evaluations!$$

Example: $(10/0.01)^5 = 10e14...$
... but function evaluations are very expensive!

What to do?

Option 3: Random search?

We can sample the space uniformly [Bergstra and Bengio 2012]



Better than grid search in various senses but still expensive to guarantee good coverage.

What to do?

Key question:

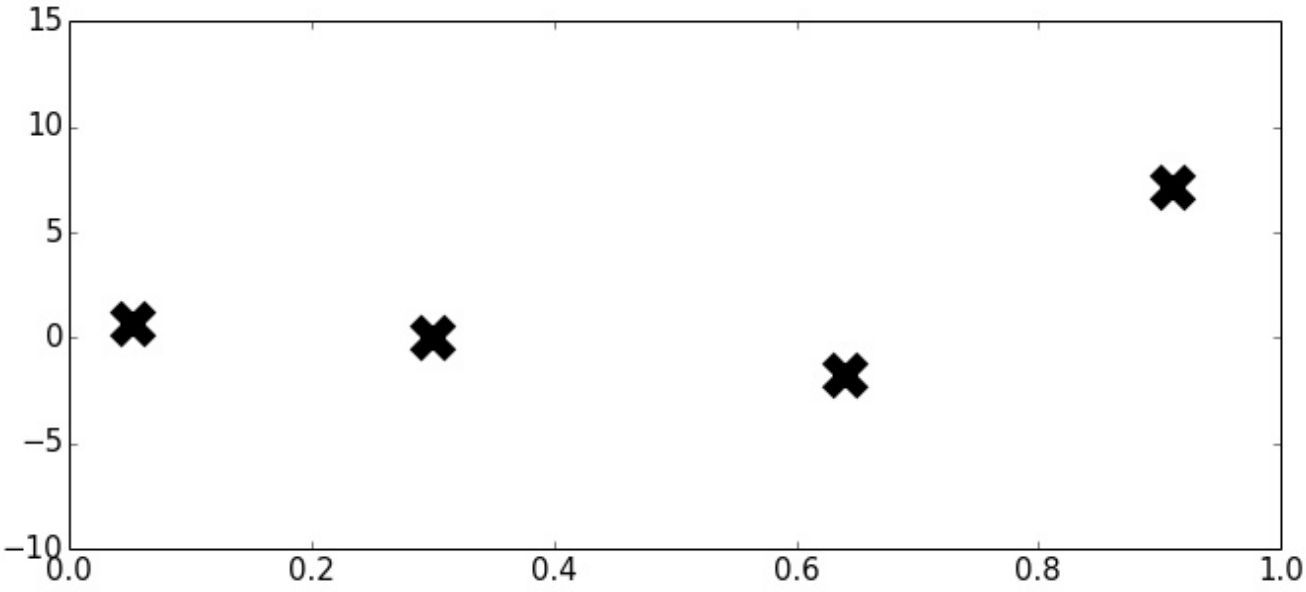
Can we do better?

Problem (the audience is encouraged to participate!)

- Find the optimum of some function f in the interval [0,1].
- f is L-Lipchitz continuous and differentiable.
- Evaluations of f are exact and we have 4 of them!

Situation

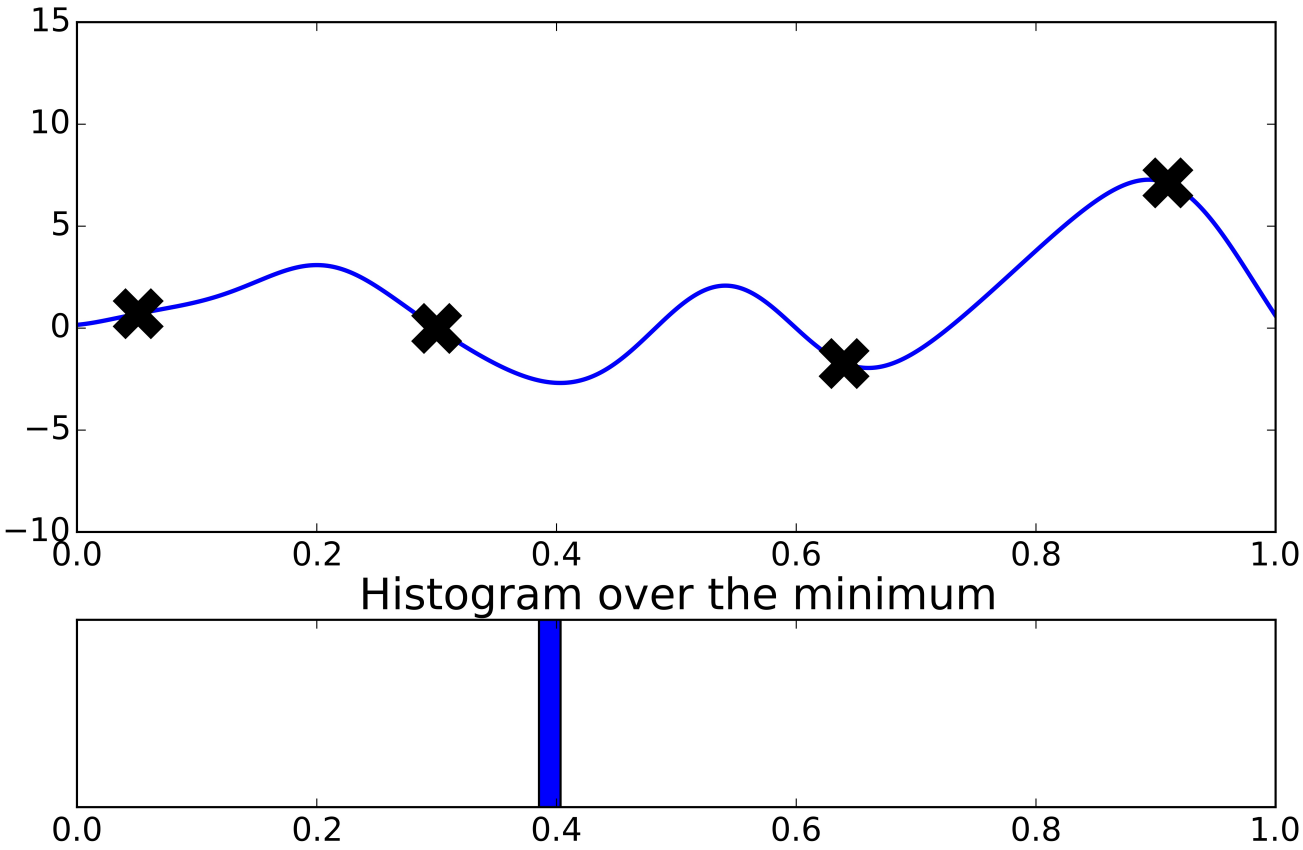
We have a few function evaluations



Where is the minimum of f?
Where should the take the next evaluation?

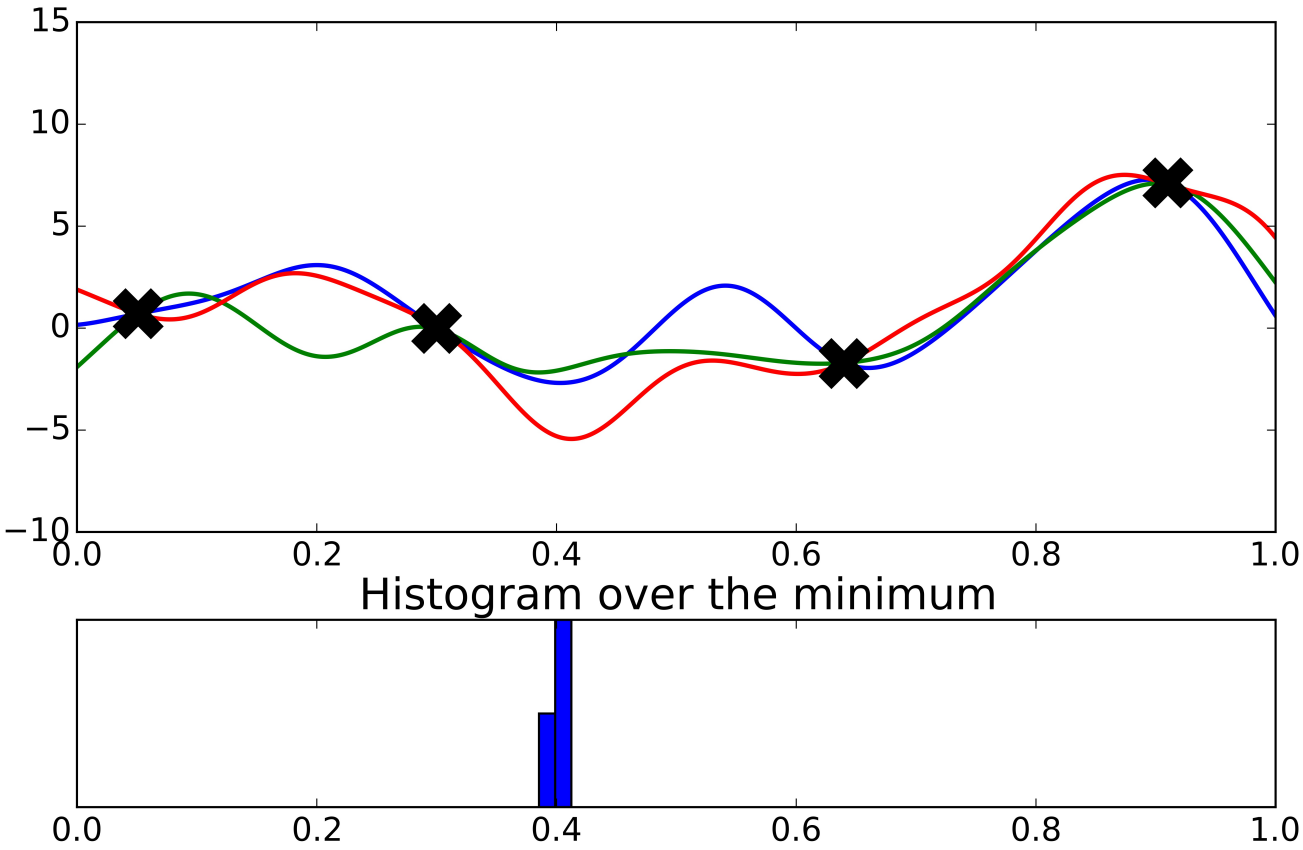
Intuitive solution

One curve



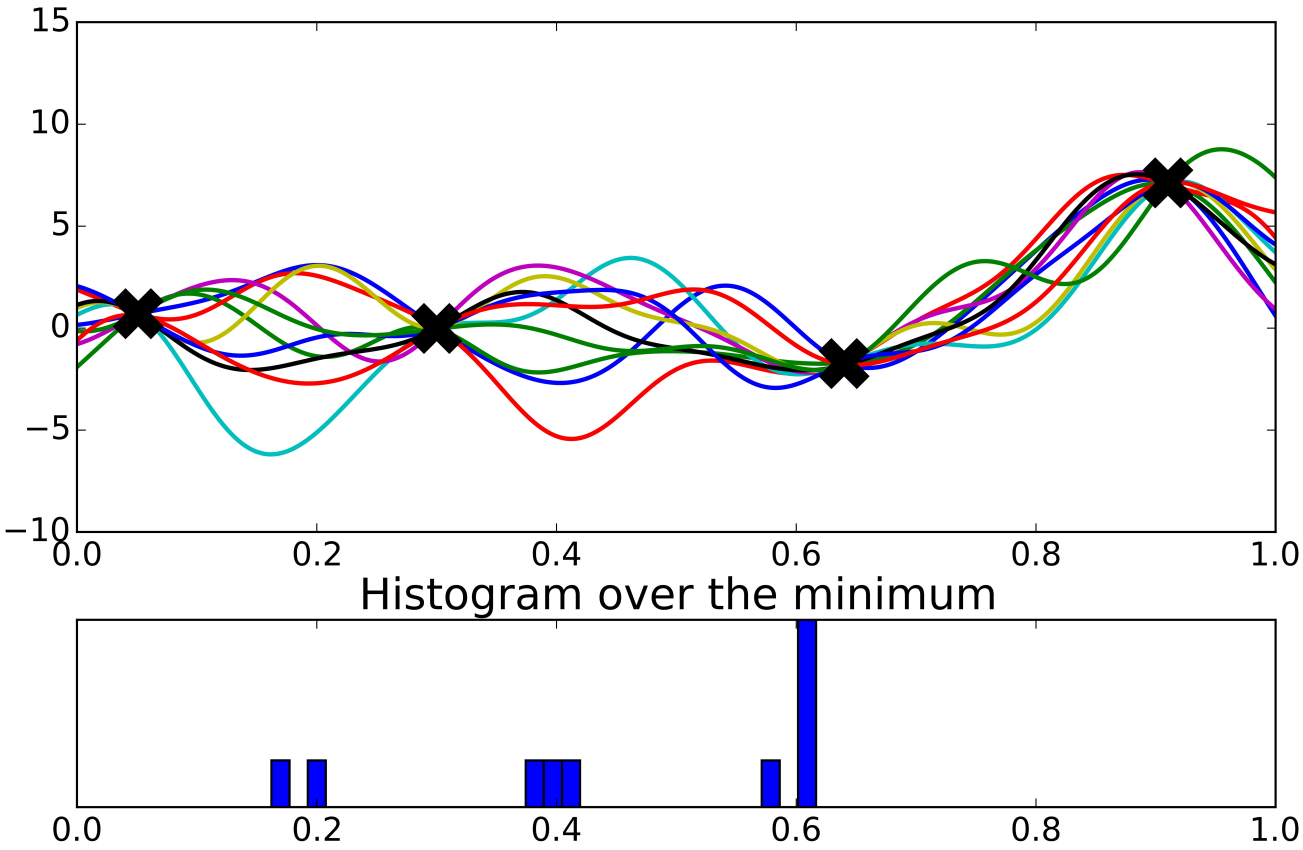
Intuitive solution

Three curves



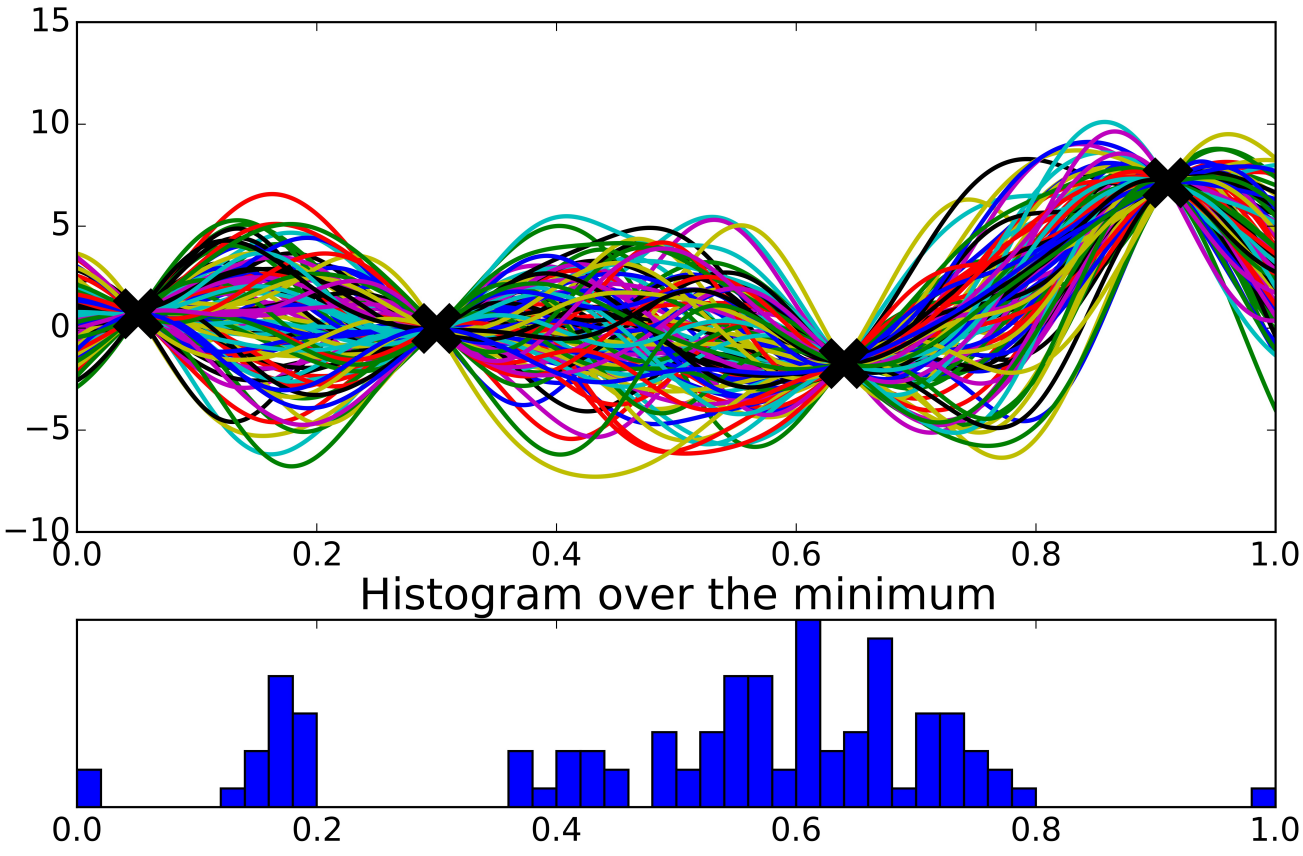
Intuitive solution

Ten curves



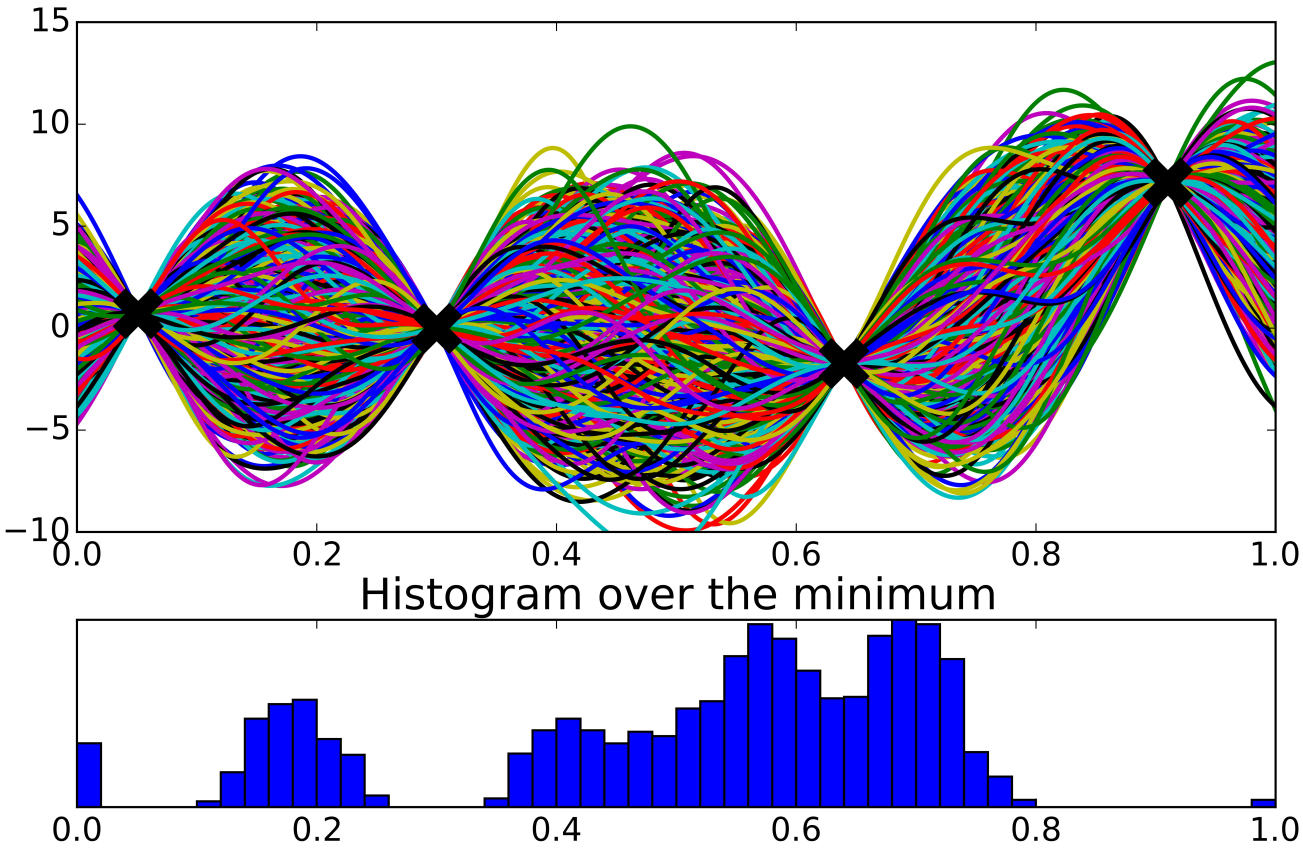
Intuitive solution

Hundred curves



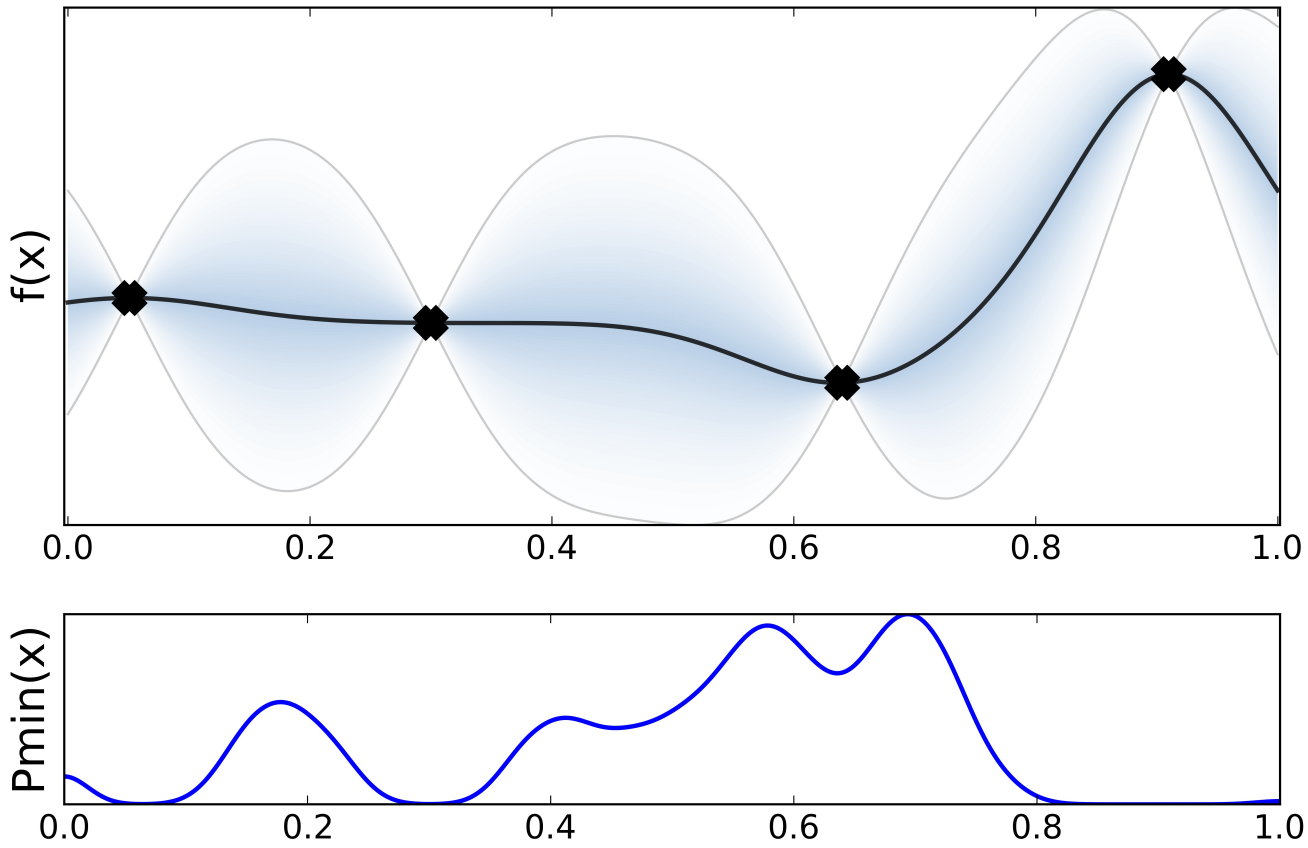
Intuitive solution

Many curves



Intuitive solution

Infinite curves



General idea: surrogate modelling

1. Use a surrogate model of f to carry out the optimization.

2. Define an utility function to collect new data points satisfying some optimality criterion: *optimization* as *decision*.

3. Study *decision* problems as *inference* using the surrogate model: use a probabilistic model able to calibrate both, epistemic and aleatoric uncertainty.

Uncertainty Quantification

Utility functions

The utility should represent our design goal:.

1. Active Learning and experimental design: Maximize the differential entropy of the posterior distribution $p(f|X, y)$ (D-optimality in experimental design).

2. Minimize the loss in a sequence $x_1, \ldots, x_n$

$$r_N = \sum_{n=1}^{N} f(x_n) - N f(x_M)$$

(1) does to a lot exploration whereas (2) encourages exploitation about the minimum of f.

Bayesian Optimisation

[Mockus, 1978]

Methodology to perform global optimisation of multimodal black-box functions.

1. Choose some *prior measure* over the space of possible objectives f.

2. Combine prior and the likelihood to get a *posterior measure* over the objective given some observations.

3. Use the posterior to decide where to take the next evaluation according to some *acquisition/loss function*.

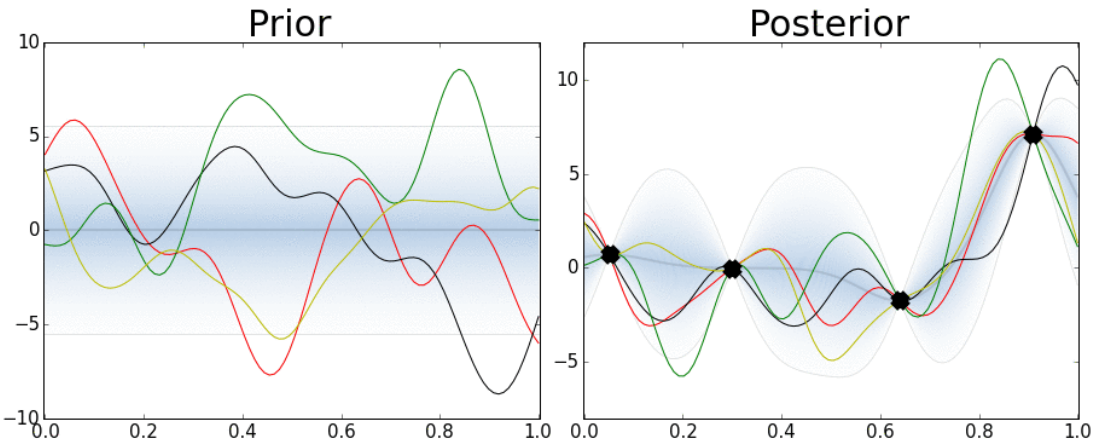
4. Augment the data.

Iterate between 2 and 4 until the evaluation budget is over.

Surrogate model: Gaussian process

Default Choice: Gaussian processes [Rasmunsen and Williams, 2006]

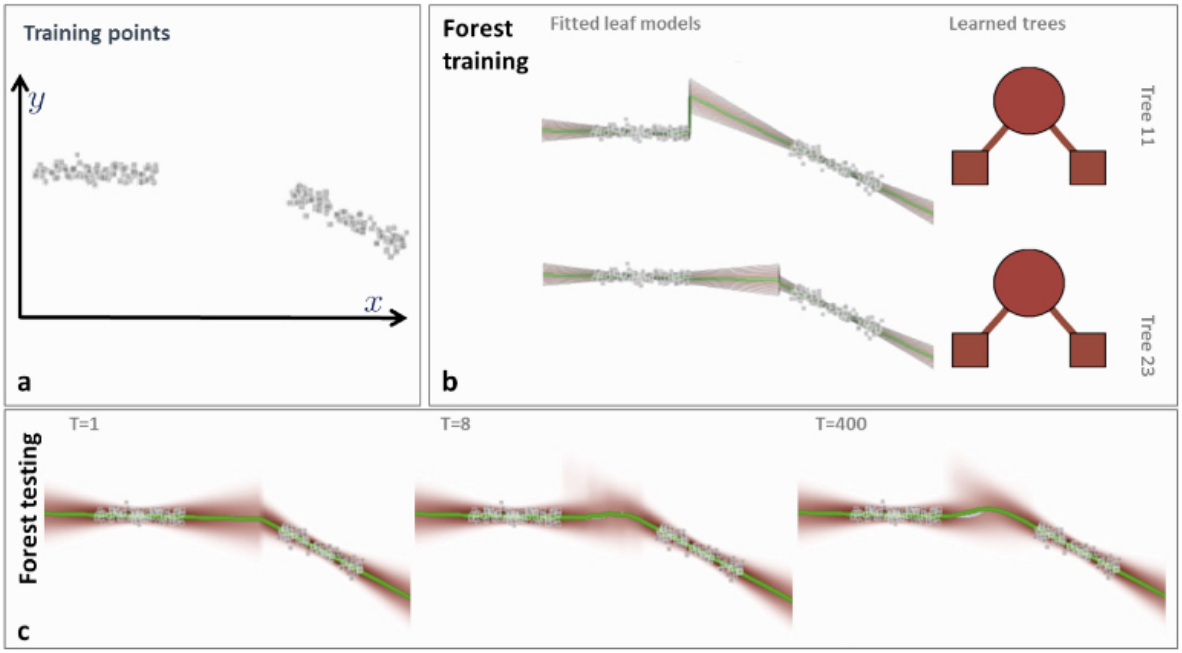
Infinite-dimensional probability density, such that each linear finite-dimensional restriction is multivariate Gaussian.



- Model $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$ is determined by the *mean function* $m(x)$ and *covariance function* $k(x, x'; \theta)$.
- Posterior mean $\mu(x; \theta, \mathcal{D})$ and variance $\sigma(x; \theta, \mathcal{D})$ can be *computed explicitly* given a dataset $\mathcal{D}$.

Other models are also possible: Random Forrest

[Criminisi et al, 2011]



Other models are also possible: t-Student processes

Student-t Processes as Alternatives to Gaussian Processes

Amar Shah
University of Cambridge

Andrew Gordon Wilson
University of Cambridge

Zoubin Ghahramani
University of Cambridge

Abstract

We investigate the Student-t process as an alternative to the Gaussian process as a non-parametric prior over functions. We derive closed form expressions for the marginal likelihood and predictive distribution of a Student-t process, by integrating away an

simple exact learning and inference procedures, and impressive empirical performances [Rasmussen, 1996], Gaussian processes as kernel machines have steadily grown in popularity over the last decade.

At the heart of every Gaussian process (GP) is a parametrized covariance kernel, which determines the properties of likely functions under a GP. Typically simple parametric kernels, such as the Gaus-

Exploration vs. exploitation

Bayesian optimization explains human active search

[Borji and Itti, 2013]

Exploration vs. exploitation

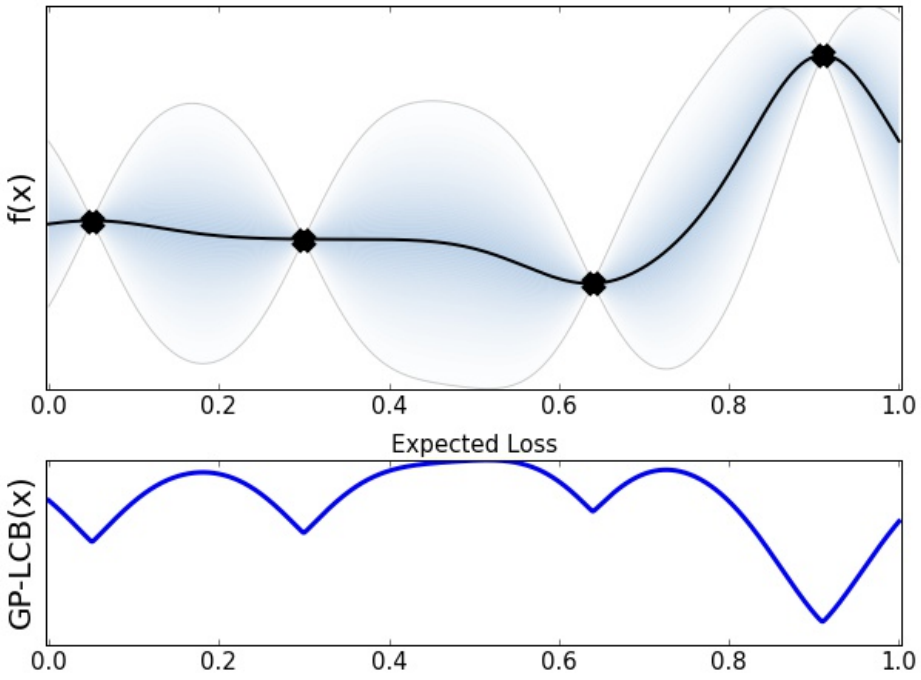
Picture source: http://peakdistrictcycleways.co.uk

GP Upper (lower) Confidence Band

[Srinivas et al., 2010]

Direct balance between exploration and exploitation:

$$\alpha_{LCB}(\mathbf{x}; \theta, \mathcal{D}) = -\mu(\mathbf{x}; \theta, \mathcal{D}) + \beta_t \sigma(\mathbf{x}; \theta, \mathcal{D})$$



GP Upper (lower) Confidence Band

[Srinivas et al., 2010]

- In noiseless cases, it is a lower bound of the function to minimize.
- This allows to computer a bound on how close we are to the minimum.
- Optimal choices available for the 'regularization parameter'.

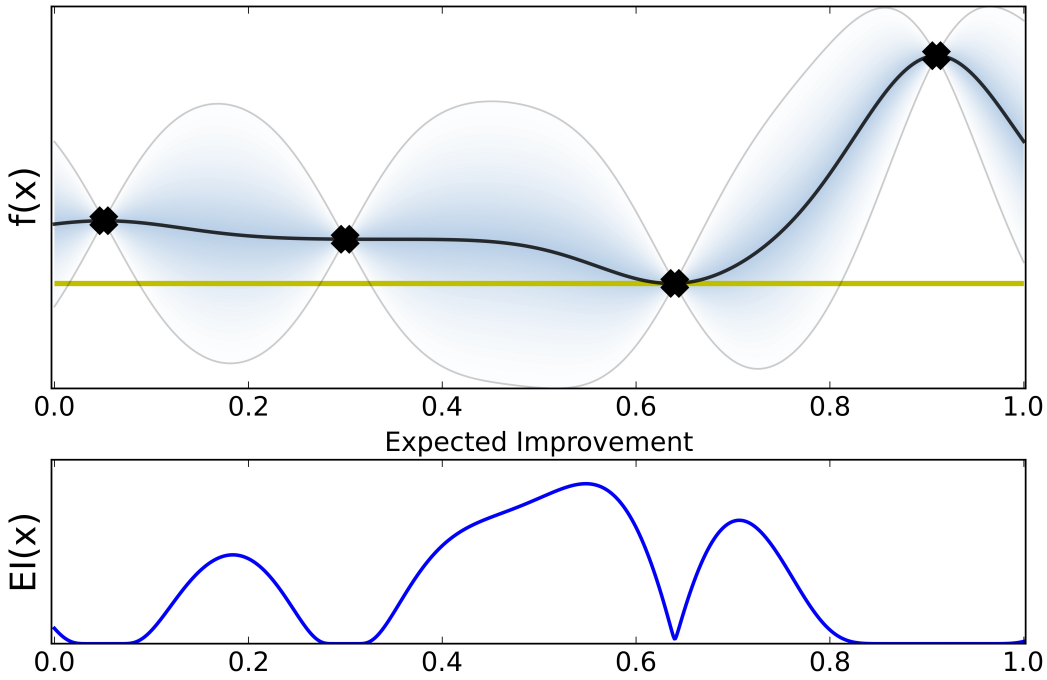
Theorem 1 *Let* $\delta \in (0,1)$ *and* $\beta_t = 2\log(|D|t^2\pi^2/6\delta)$. *Running GP-UCB with* β_t *for a sample* f *of a GP with mean function zero and covariance function* $k(\boldsymbol{x}, \boldsymbol{x}')$, *we obtain a regret bound of* $\mathcal{O}^*(\sqrt{T\gamma_T \log|D|})$ *with high probability. Precisely, with* $C_1 = 8/\log(1+\sigma^{-2})$ *we have*

$$\Pr\left\{R_T \leq \sqrt{C_1 T \beta_T \gamma_T} \quad \forall T \geq 1\right\} \geq 1-\delta.$$

Expected Improvement

[Jones et al., 1998]

$$\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$$



Expected Improvement

[Jones et al., 1998]

- Perhaps the most used acquisition.
- Explicit for available for Gaussian posteriors.
- It is too greedy in some problems. It is possible to make more explorative adding a 'explorative' parameter

$$\alpha_{EI}(\mathbf{x};\theta,\mathcal{D}) = \sigma(\mathbf{x};\theta,\mathcal{D})(\gamma(x)\Phi(\gamma(x))) + \mathcal{N}(\gamma(x);0,1).$$

where

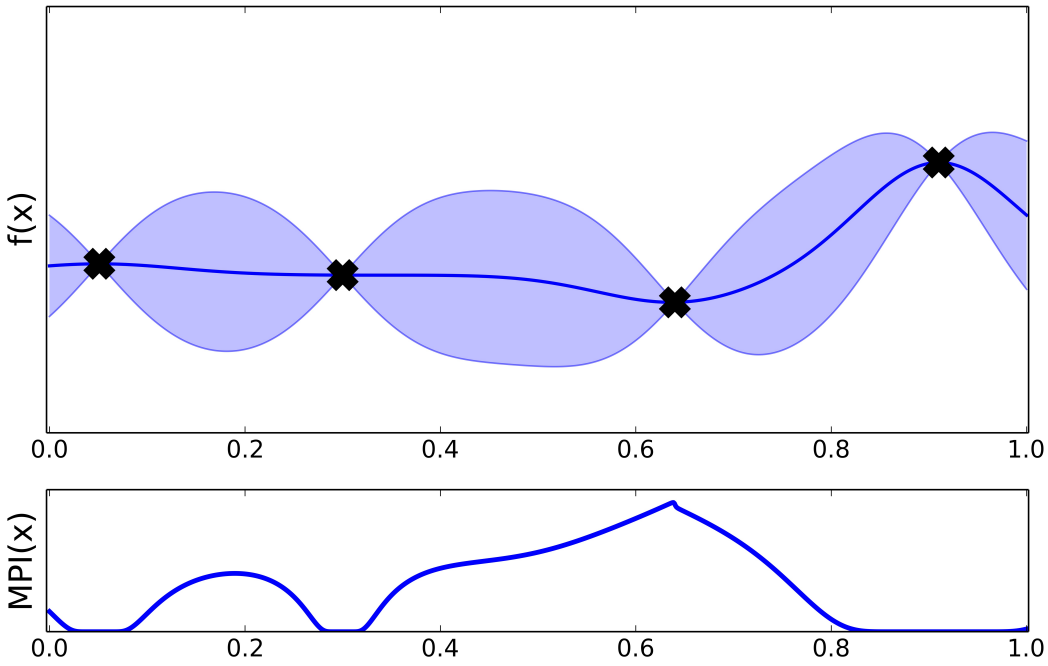
$$\gamma(x) = \frac{f(x_{best}) - \mu(\mathbf{x};\theta,\mathcal{D}) + \psi}{\sigma(\mathbf{x};\theta,\mathcal{D})}.$$

Maximum Probability of Improvement

[Hushner, 1964]

$$\gamma(\mathbf{x}) = \sigma(\mathbf{x}; \theta, \mathcal{D})^{-1}(\mu(\mathbf{x}; \theta, \mathcal{D}) - y_{best})$$

$$\alpha_{MPI}(\mathbf{x}; \theta, \mathcal{D}) = p(f(\mathbf{x}) < y_{best}) = \Phi(\gamma(\mathbf{x}))$$



Maximum Probability of Improvement

[Hushner, 1964]

- First used acquisition: very intuitive.
- Less used in practice.
- Explicit for available for Gaussian posteriors.

$$\alpha_{MPI}(\mathbf{x}; \theta, \mathcal{D}) = \Phi(\gamma(x))).$$

where

$$\gamma(x) = \frac{f(x_{best}) - \mu(\mathbf{x}; \theta, \mathcal{D}) + \psi}{\sigma(\mathbf{x}; \theta, \mathcal{D})}.$$

Information-theoretic approaches

[Hennig and Schuler, 2013; Hernández-Lobato et al., 2014]

$$\alpha_{ES}(\mathbf{x}; \theta, \mathcal{D}) = H[p(x_{min}|\mathcal{D})] - \mathbb{E}_{p(y|\mathcal{D},\mathbf{x})}[H[p(x_{min}|\mathcal{D} \cup \{\mathbf{x}, y\})]]$$